ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 4, Nomor 3, Tahun 2015, Halaman 651-659
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

KLASIFIKASI TINGKAT KELANCARAN NASABAH DALAM MEMBAYAR PREMI DENGAN MENGGUNAKAN METODE REGRESI LOGISTIK ORDINAL DAN NAÏVE BAYES
(Studi Kasus pada Asuransi AJB Bumiputera Tanjung Karang Lampung)

Ria Sutitis[1], Suparti[2], Dwi Ispriyanti[3]
[1]Mahasiswa Jurusan Statistika FSM Universitas Diponegoro
[2,3]Staff Pengajar Jurusan Statistika FSM Universitas Diponegoro
e-mail: sutitis81@gmail.com

ABSTRACT

In the insurance companies a problem that often arises is the amount of customer debt in paying premiums, so it needs a system that can classify customers in the group not well, less smoothly, and smooth in paying premiums. Used two methods to perform the classification of payment premium status which is Regression Logistics Ordinal and Naïve Bayes. Variables used in determining whether a payment premium status are gender, marital status, age, work, income, insurance period, and the payment of premium. In Regression Logistics Ordinal, significant variables to the model are gender, marital status, age, insurance period, and the payment of premium. For significant variables used in the classification. Payment premium status of the data processing methods of Regression Logistics Ordinal with accuracy obtained is equal to 50.90% and the Naïve Bayes method obtained is equal to 55.41%. Based on the level of accuracy, the classification of data payment premium status of insurance AJB Bumiputera Tanjung Karang Lampung using the Naïve Bayes method has a greater degree of accuracy than the Regression Logistics Ordinal method.

Keywords: Payment Premium Status, Classification, Naïve Bayes, Regression Logistics Ordinal

1. PENDAHULUAN

1.1. Latar Belakang

Premi merupakan pendapatan bagi perusahaan asuransi, yang jumlahnya ditentukan dalam suatu persentase atau tarif tertentu dari jumlah yang dipertanggungkan. Pendapatan premi untuk perusahaan asuransi ditentukan oleh jumlah premi yang dibayar oleh nasabah. Permasalahan yang sering timbul dalam perusahaan asuransi adalah banyaknya nasabah yang menunggak dalam membayar premi, sehingga diperlukan sebuah sistem yang dapat mengklasifikasikan nasabah yang masuk ke dalam kelompok lancar, kelompok kurang lancar dan kelompok tidak lancar dalam membayar premi.

Dalam membayar premi yang dapat diklasifikasikan ke dalam kelompok lancar, kelompok kurang lancar, dan kelompok tidak lancar dapat disebabkan karena beberapa hal. Data dari faktor-faktor atau hal-hal yang mempengaruhi pembayaran premi berbentuk data katagorik. Dengan data yang berbentuk katagorik, maka dalam menganalisis dapat menggunakan Analisis Data Kategorik. Dalam penelitian ini variabel respon berbentuk kategorik bertingkat atau ordinal, sehingga dalam laporan ini menggunakan Metode Regresi Logistik Ordinal.

Sebuah perusahaan asuransi pastilah mempunyai data yang begitu besar. Banyak yang belum menyadari bahwa dari pengolahan data – data tersebut dapat memberikan informasi berupa klasifikasi data nasabah yang akan bergabung pada perusahaan itu sendiri. Penggunaaan teknik data mining diharapkan mampu memberikan informasi yang berguna tentang teknik klasifikasi data nasabah yang akan bergabung dalam kelompok mana dalam membayar premi. Salah satu teknik data mining yang dapat digunakan adalah metode Klasifikasi Naïve Bayes.

1.2. Rumusan Masalah

Perbandingan ketepatan klasifikasi pembayaran premi nasabah asuransi AJB Bumiputera Tanjung Karang Lampung dengan menggunakan metode Regresi Logistik Ordinal dan metode Naïve Bayes. Penelitian mengenai status pembayaran premi nasabah asuransi AJB Bumiputera ini dikhususkan untuk nasabah asuransi AJB Bumiputera Tanjung Karang Lampung sesuai dengan rekapitulasi pada tahun 2014.

1.3. Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Menghitung ketepatan klasifikasi dengan menggunakan metode Regresi Logistik Ordinal.
2. Menghitung ketepatan klasifikasi dengan menggunakan metode Naïve Bayes.
3. Memilih ketepatan klasifikasi terbaik yang dihasilkan antara metode Regresi Logistik Ordinal dan metode Naïve Bayes.

2. TINJAUAN PUSTAKA

2.1. Regresi Logistik Ordinal

Model logistik untuk data respon ordinal sering disebut pula dengan model logit kumulatif. Logit kumulatif digunakan untuk menganalisis hubungan antara variabel dependen yang berupa *polytomous ordinal response* dengan sekumpulan variabel independen, yang dapat disebut sebagai *factors or covariates*. Rancangan untuk regresi ordinal didasarkan pada Methodology of McCullgh (1980).

Respon dalam model logit kumulatif dapat berupa data bertingkat yang diwakili dengan angka 1,2,3,…,k dimana k adalah banyaknya kategori respon. Bentuk model logit kumulatif untuk respon ordinal dengan k kategori yaitu:

$$Logit[C_j] = log\left[\frac{C_j}{1-C_j}\right] = \theta_j - \boldsymbol{\beta}^T\mathbf{X} \quad (1)$$

dengan

$$C_j = [P(Y \leq j)] = \pi_j = \frac{e^{(\theta_j - \boldsymbol{\beta}^T\mathbf{X})}}{1+e^{(\theta_j - \boldsymbol{\beta}^T\mathbf{X})}} \quad (2)$$

π_j = peluang kategori respon ke-j, θ_j = konstanta (j = 1,2,…,k-1), $\boldsymbol{\beta}$ = vektor parameter koefisien ($\beta_1, \beta_2, \ldots, \beta_p$), $\mathbf{X}$ = vektor variabel bebas ($X_1, X_2, \ldots, X_p$); p adalah banyaknya variabel bebas.

A. Uji Signifikansi

1. Uji Parameter secara Keseluruhan

Dalam Hosmer (1989) uji parameter secara keseluruhan atau uji rasio likelihood diperoleh dengan cara membandingkan fungsi log likelihood dari seluruh variabel bebas dengan fungsi log likelihood tanpa variabel bebas.

a. Hipotesis

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1 : paling sedikit salah satu dari $\beta_r \neq 0$ dengan r = 1,2,…,p

b. Statistik Uji Rasio Likelihood

X^2_{hit} = -2 log $\left(\frac{likelihood\ tanpa\ variabel\ bebas}{likelihood\ dengan\ variabel\ bebas}\right)$

= 2 log [likelihood dengan variabel bebas] – 2 log [likelihood tanpa variabel bebas]

c. Kriteria Uji

Tolak H_0 jika $X^2_{hit} > X^2_{(\alpha,p)}$

2. Uji Parameter secara Individu

Dalam Hosmer (1989) uji parameter secara individu diperoleh dengan cara menguadratkan rasio estimasi parameter dengan estimasi standar errornya. Uji ini menggunakan uji wald yang berfungsi menguji signifikansi tiap parameter.

a. Hipotesis

$H_0 : \beta_r = 0$

$H_1 : \beta_r \neq 0$, dengan r = 1,2,…,p

b. Statistik Uji:

$$W_r = \left[\frac{\hat{\beta}_r}{\widehat{SE}(\hat{\beta}_r)}\right]^2$$

c. Kriteria Uji

Tolak H_0 jika $W_r > \chi^2_{(\alpha,1)}$

2.2. Klasifikasi Naïve Bayes

Klasifikasi adalah proses pencarian sekumpulan model atau fungsi yang menggambarkan dan membedakan kelas data dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu objek yang belum diketahui kelasnya (Han and Kamber, 2006).

Pada klasifikasi Naïve Bayes, proses pembelajaran lebih ditekankan pada mengestimasi probabilitas. Keuntungan dari pendekatan ini yaitu pengklasifikasian akan mendapatkan nilai error yang lebih kecil ketika data set berjumlah besar (Berry, 2006). Selain itu menurut Han and Kamber (2006) klasifikasi Naïve Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam basis data dengan jumlah yang besar.

Formula Naïve Bayes untuk klasifikasi menurut Prasetyo (2012) yaitu:

$$P(Y|X) = \frac{P(Y)\Pi_{i=1}^{q} P(X_i|Y)}{P(X)} \quad (3)$$

Dimana:

$P(Y|X)$ = probabilitas data dengan vektor X pada kelas Y

$P(Y)$ = probabilitas awal kelas Y (*prior probability*)

$\Pi_{i=1}^{q} P(X_i|Y)$ = probabilitas independen kelas Y dari semua fitur dalam vektor X.

$P(X)$ = probabilitas X

Probabilitas P(X) selalu tetap sehingga dalam perhitungan prediksi nantinya dapat diabaikan dan hanya menghitung bagian $P(Y)\Pi_{i=1}^{q} P(X_i|Y)$ saja dengan memilih nilai yang terbesar sebagai kelas hasil prediksi.

Untuk fitur dengan tipe numerik (kontinu) ada perlakuan khusus sebelum diproses menggunkan metode Naïve Bayes yaitu mengasumsikan bentuk distribusi tertentu misalkan saja fitur berdistribusi Gaussian untuk mempresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas $P(X_i|Y)$, sedangkan distribusi Gaussian sendiri

dikarakteristikkan dengan parameter yaitu mean (μ) dan varian (σ^2). Untuk setiap kelas y_j, probabilitas bersyarat kelas y_j untuk fitur X_i adalah:

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} \exp - \frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}$$

Parameter μ_{ij} bisa di dapat dari mean sampel $X_i(\bar{X})$ dari semua data training yang menjadi milik kelas y_j, sedangkan σ_{ij}^2 dapat diperkirakan dari varian sampel (s^2) dari data training.

2.3. Teknik Validasi Model

Cross-validasi (*cross validation*) atau yang sering disebut dengan estimasi rotasi merupakan teknik komposisi dalam penentuan banyaknya data training dan data testing yang akan digunakan. Dalam cross-validasi metode yang dapat digunakan salah satunya adalah metode Holdout.

Dalam metode holdout, data awal yang diberi label dipartisi ke dalam dua himpunan secara random yang dinamakan data training dan data testing. Model klasifikasi selanjutnya dihasilkan dari data training dan kinerjanya dievaluasi pada data testing. Proporsi data yang dicadangkan untuk data training dan data testing tergantung pada analisis misalnya 50% : 50% atau 2/3 untuk training dan 1/3 untuk testing, namun menurut Witten (2005) serta Han and Kamber (2006) pada umumnya perbandingan yang digunakan yaitu 2:1 untuk training berbanding data testing.

2.4. Evaluasi Ketepatan Hasil Klasifikasi

Untuk mengetahui hasil klasifikasi sudah akurat atau belum, maka dilakukan uji ketepatan hasil evaluasi. Untuk mengetahui ketepatan tersebut dapat dilakukan dengan menggunakan APER (*Apparent Error Rate*).

APER (*Apparent Error Rate*) atau yang disebut laju error merupakan ukuran evaluasi yang digunakan untuk melihat peluang kesalahan klasifikasi yang dihasilkan oleh suatu fungsi klasifikasi. Semakin kecil nilai APER maka hasil pengklaifikasian semakin baik (Prasetyo, 2012).

Formulasi untuk menghitung APER (Johnson and Wichern, 2007) yaitu:

$$\text{APER} = \frac{f_{12}+f_{13}+f_{21}+f_{23}+f_{31}+f_{32}}{f_{11}+f_{12}+f_{13}+f_{21}+f_{22}+f_{23}+f_{31}+f_{32}+f_{33}} \text{ x } 100\%$$

3. METODE PENELITIAN

3.1. Sumber Data

Data yang akan digunakan sebagai studi kasus pada tugas akhir ini berupa data sekunder yang diambil dari data nasabah asuransi untuk AJB Bumiputera Tanjung Karang Lampung. Data nasabah yang digunakan yaitu sebanyak 1778.

3.2. Variabel Data

Variabel yang diklasifikasikan untuk metode Naïve Bayes dan sebagai variabel respon pada metode Regresi Logistik Ordinal adalah status pembayaran premi yang diukur dengan skala ordinal dengan tiga kategori yaitu tidak lancar, kurang lancar, dan lancar. Sedangkan variabel penentu yang digunakan dalam mengklasifikasikan data nasabah yaitu:

a. Jenis Kelamin

   Variabel jenis kelamin nasabah dikelompokkan dalam dua kategori yaitu laki-laki dan perempuan.

b. Status Perkawinan

   Variabel status perkawinan dikelompokkan dalam dua kategori yaitu bersuami/istri dan tidak bersuami/istri.

c. Usia

   Variabel usia dikelompokkan dalam tiga kategori yaitu 20-29 Tahun, 30-40 Tahun, dan > 40 Tahun.

d. Pekerjaan

   Variabel pekerjaan dikelompokkan dalam tiga kategori yaitu PNS, pegawai swasta, dan wiraswasta.

e. Penghasilan

   Variabel penghasilan dikelompokkan dalam tiga kategori yaitu < 25 juta pertahun, 25-50 juta pertahun, dan > 50 juta pertahun.

f. Masa Pembayaran Premi

   Variabel masa pembayaran asuransi dikelompokkan dalam tiga kategori yaitu 5-10 tahun, 11-15 tahun, dan > 15 tahun.

g. Cara Pembayaran Premi

   Variabel cara pembayaran premi dikelompokkan dalam tiga kategori yaitu triwulan, semesteran, dan tahunan

3.3. Metode Analisis Data

Data sekunder yang terkumpul dianalisis dan diolah menggunakan metode Regresi Logistik Ordinal dan metode Naïve Bayes. Dimana prosedur pengolahannya adalah:

1. Membagi data menjadi data training dan data testing
2. Membentuk model dari regresi logistic ordinal
3. Menguji signifikasi parameter secara keseluruhan dan secara individu.
4. Menghitung nilai peluang dengan menggunakan data testing.
5. Pada Regresi Logistik Ordinal, kelas hasil prediksi adalah kelas yang memiliki nilai peluang paling tinggi.
6. Untuk klasifikasi menggunakan Naïve Bayes, variabel yang digunakan adalah variabel yang signifikan terhadap model pada Regresi Logistik Ordinal.
7. Menghitung probabilitas awal (*prior probability*) Y (P(Y))
8. Menghitung nilai probabilitas independen kelas Y dari semua fitur dalam vektor X ($\Pi^{q}_{i=1}$ P(X_i|Y))
9. Menghitung *posterior probabiliy* untuk masing-masing klasifikasi (P(Y|X)
10. Hasil prediksi yang didapat adalah nilai maksimum dari hasil penghitungan *posterior probability*.

4. HASIL DAN PEMBAHASAN

4.1. Regresi Logistik Ordinal

Dengan menggunakan perbandingan 75%:25% data training berbanding data testing, pada Regresi Logistik Ordinal didapat model sebagai berikut:

Logit 1 (Tidak Lancar) = – 0,104 – 0,313 JK(1) – 0,703 Stat_Kawin(1) – 0,353 Usia(1) – 0,055 Usia(2) – 0,909 MA(1) – 0,794 MA(2) + 0,382 Pembayaran(1) + 0,458 Pembayaran(2)

Logit 2 (Kurang Lancar)= 1,146 – 0,313 JK(1) – 0,703 Stat_Kawin(1) – 0,353 Usia(1) – 0,055 Usia(2) – 0,909 MA(1) – 0,794 MA(2) + 0,382 Pembayaran(1) + 0,458 Pembayaran(2)

Dengan model diatas, dihitung nilai C_j menggunakan data testing, kemudian dihitung peluang untuk masing-masing kelas. Untuk hasil prediksi merupakan nilai peluang yang paling besar. Untuk ketepatan klasifikasi data testing menggunakan Regresi Logistik Ordinal, dapat menggunakan matriks konfusi pada Tabel 1.

Tabel 1. Matriks Konfusi Regresi Logistik Ordinal

f_{ij}		Kelas Prediksi		
		Kelas = 1	Kelas = 2	Kelas = 3
Kelas Asli	Kelas = 1	12	0	89
	Kelas = 2	5	7	105
	Kelas = 3	9	10	207

Diperoleh nilai APER dan akurasi sebagai berikut:

APER = 218/444 = 0,4910 = 49,10%

Akurasi = 1- APER = 1 – 0,4910 = 0,5090 = 50,90%

4.2. Naïve Bayes

Pengklasifikasian menggunakan metode Naïve Bayes diperoleh hasil prediksi sebagai berikut:

Tabel 2. Matriks Konfusi Naïve Bayes

f_{ij}		Kelas Prediksi		
		Kelas = 1	Kelas = 2	Kelas = 3
Kelas Asli	Kelas = 1	3	12	86
	Kelas = 2	4	13	86
	Kelas = 3	1	9	230

Diperoleh nilai APER dan Akurasi sebagai berikut:

APER = 198/444 = 0,4459 = 44,59%

Akurasi = 1 – APER = 1 – 0,4459 = 0,5541 = 55,41%

4.3. Pemilihan Ketepatan Klasifikasi Terbaik

Pemilihan ketepatan klasifikasi terbaik berdasarkan akurasi terbesar. Perbandingan ketepatan klasifikasi Regresi Logistik Ordinal dan Naïve Bayes dapat dilihat pada tabel 3.

Tabel 3. Perbandingan Ketepatan Klasifikasi

	Akurasi	Laju Error
Regresi Logistik Ordinal	50,90%	49,10%
Naïve Bayes	55,41%	44,59%

Pada Tabel 3 dapat dilihat bahwa nilai akurasi ketepatan klasifikasi Naïve Bayes lebih besar dibandingkan dengan nilai akurasi ketepatan klasifikasi Regresi Logistik Ordinal, sehingga dapat disimpulkan bahwa ketepatan klasifikasi terbaik adalah klasifikasi Naïve Bayes dengan nilai akurasi sebesar 55,41%.

5. KESIMPULAN

Dari hasil dan analisis data dapat disimpulkan:

1. Pada Regresi Logistik Ordinal variabel bebas yang signifikan terhadap model adalah variabel status kawin, usia, masa pembayaran, dan cara pembayaran premi. Dengan 444 data testing didapat laju error sebesar 49,10% dan akurasi ketepatan klasifikasi adalah sebesar 50,90%
2. Klasifikasi Naïve Bayes dengan data testing sebesar 444 didapat laju error sebesar 44,59% dan akurasi ketepatan klasifikasi sebesar 55,41%.
3. Ketepatan klasifikasi terbaik adalah berdasarkan nilai akurasi terbesar, dalam permasalahan ini klasifikasi terbaik adalah klasifikasi Naïve Bayes.

DAFTAR PUSTAKA

Han.J and Kamber, M. 2006. Data Mining Concepts and Techniques, Second Edition. California: Morgan Kaufman

Hosmer, D.W. and Lemeshow, S. (1989), *Applied Logistic Regression,* John Wiley and Sons Inc, Canada.

McCullagh, P. 1980. Regression Models for Ordinal data. Jornal of the Royal Statistical Sociery. Series B (Methodological), Volume 42, Issue 2 (1980, 109-142)

Prasetyo, E. 2006. Data Mining Konsep dan Aplikasi Menggunakan MATLAB. Yogyakarta: ANDI Yogyakarta

Witten, I.H. and Frank, E. 2005. Data Mining Pratical Machine Learning Tools and Teachniques, Second Edition. California: Morgan Kaufman